DOCUMENT RESUME

ED 039 747 56 EM 008 082

AUTHOR Carroll, John B.
TITLE Measurement and Evaluation in Educational Technology.
INSTITUTION Academy for Educational Development, Inc., Washington, D.C.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO BR-8-0571
PUB DATE [70]
NOTE 26p.; This is one of the support papers for "To Improve Learning: a Report to the President and the Congress of the United States by the Commission on Instructional Technology", ED 034 905

EDRS PRICE EDRS Price MF-$0.25 HC-$1.40
DESCRIPTORS Effective Teaching, *Evaluation, *Instructional Technology, *Measurement

ABSTRACT

Educational measurement and evaluation are technologies which are central to the operation and improvement of the educational process, because they enable the educator to know crucially important things about pupil characteristics and achievements. It also furnishes him with a valid basis for judging the worth and effectiveness of educational programs and innovations, improving them in both broad and detailed features. There is still a large gap between what it is possible to accomplish through measurement and evaluation and what has actually been accomplished. This gap can be filled by training more research and development specialists, training teachers and administrators to utilize research and development results more effectively, and providing adequate funds for these training, research, and development activities. (Author/GO)

Measurement and Evaluation in Educational Technology

by John B. Carroll*

Measurement and evaluation have long played, and will continue to play, a major role in the development of educational technology. This paper will first point out that educational measurement and evaluation is itself a technology: it will then proceed to describe how this technology has been applied, and can be even better applied than in the past, to the development and utilization of instructional procedures and materials, particularly those using newer technologies whereby the interaction between learner and content to be learned can be controlled and monitored more efficiently than in traditional classroom instruction.

Educational Measurement and Evaluation as a Technology

Educational measurement is a technology in the sense that it consists of a set of procedures and developed products founded on mathematical principles and scientific concepts.

At the base of this technology are the theories and formulations of mathematical statistics, which yield methods of collecting, summarizing, and interpreting both quantitative and qualitative data, particularly data that exhibit variation over populations or over samples of populations with respect to given characteristics. The research worker in educational measurement is required to be thoroughly familiar with such statistical techniques as multivariate correlational analysis, factor analysis, analysis

---

* John B. Carroll is senior research psychologist at Educational Testing Service, Princeton, New Jersey.

of variance and covariance, tests of statistical significance, survey sampling methods, and the design of experiments.

Another discipline that is fundamental to educational measurement is psychology. Psychology provides educational measurement with basic information on the characteristics that differentiate individuals and on the processes of maturation and learning that are involved in changes in skill, knowledge, and performance. Indeed, a theory of individual differences (Anastasi, 1958) underlies all work in educational measurement and evaluation. This is so because educational measurement must take into account the status of the learner before he starts to learn a particular task or course content and also the processes of learning and motivation that come into play in behavioral changes.

A special discipline or field of inquiry that depends both on mathematical statistics and psychology is what has been called "test theory." Test theory is a theory of measurement as applied to the kinds of measurements that are used in psychology and education. As developed to a high degree of technical adequacy and sophistication by such writers as Lord and Novick (1968), it specifies methods whereby the reliability (accuracy of measurement in the sense of freedom from error) and validity (meaningfulness and predictive efficacy) of measuring procedures can be evaluated and/or improved.

Among the technological products that have been developed within the field of educational measurement are large numbers of standardized tests for measuring various aspects of intelligence, personality, vocational interests, social attitudes, and educational achievements (Buros, 1965). But almost of equal importance in educational evaluation are the instruments

that can be, and are, constructed by teachers and research workers for the measurement of particular traits or achievements. To be sure, not all these measurement instruments have satisfactory reliability and validity for the purposes for which they are intended, but it remains true that a well developed theory of measurement is available for the design and evaluation of any particular measurement device or procedure.

Other technological products of educational measurement include standard experimental designs (Campbell & Stanley, 1963), computer programs, and special machines for scoring test answer sheets. The very extensive research literature can also be considered as a technological outcome of educational measurement and evaluation (Harris, 1960; Gage, 1963).

Definitions of Measurement and Evaluation

The ordinary meaning of measurement is fairly well understood. One measures some object or entity, with respect to a given characteristic or trait, by some operation that assigns that object a value on a scale. The scale may be purely nominal, consisting simply of an unordered series of categories, or it may be a quantitative scale in which successive values are at least ordered in magnitude. The units of some scales may have still other properties such as equality and additivity.

For example, one may classify or measure a person with respect to sex (where "male" and "female" represent two points on a nominal scale), scholastic rank in class (where the scale is merely ordinal), "intelligence" (where the units of the scale are approximately equal), or weight (where the units are not only equal but also additive).

Few educational measurements are based on scales with additive units, but many of them have scales whose units can be regarded as approximately equal; such scales are known as interval scales. The errors of measurement are frequently quite large, however, in comparison to those usually encountered in the physical sciences. Also, educational measurements are sometimes of questionable validity, in the sense that it is not certain exactly what is being measured. It is the task of technology in educational measurement to fashion measuring procedures that are as free from error and vagueness as possible.

Evaluation--the rendering of a value judgment--goes beyond measurement. It may utilize measurements as data entering into the judgmental process, but it depends more importantly upon the use of standards and criteria.

A simple kind of evaluation occurs when one interprets the result of an educational test. If one asks whether a given score is "average," "excellent," or "poor," with respect to a representative group of test-takers, the interpretation may be said to be norm-referenced. If one can interpret a score as reflecting a certain distinct range of behaviors or a specific degree of mastery of subject-matter, we may say that the evaluation is criterion-referenced.

In a broader context, however, evaluation refers to the assessment of educational programs and their components with respect to the extent to which they achieve their stated goals and with respect to the cost (in time, money, effort, or inconvenience) of achieving these goals. It considers the degree to which the program fosters or retards student progress, whether in subject-matter skills and knowledges or in the formation of desirable interests, attitudes, and personality traits. Evaluation may even

extend to the assessment of the worthwhileness of the stated goals of a program, but such assessment must be made more with reference to a philosophy of education than with reference to technological criteria.

Evaluating Educational Programs and Their Components

Educational programs (or their components, such as curricula, textbooks, films, etc.) can be evaluated as final products, with a view to final acceptance or rejection. This is the traditional view of evaluation. Recently, however, it has come to be realized that an equally important kind of evaluation can be done in the course of developing a program, with a view to modifying and shaping it to yield best results. In the terminology introduced by Scriven (1967), the former type of evaluation is "summative" while the latter is "formative."

The work of evaluation, whether it is "formative" or "summative," begins with the attempt to state the objectives of the educational procedure or product being investigated, that is, to state in detail what kinds of changes in skill, knowledge, or performance are desired in learners. Further, it is important to include in the statement of objectives information on what kinds of learners these changes are desired in--their characteristics in terms of age, intellectual maturity, prior learning experiences, and (sometimes) personality.

The task of stating educational objectives is not as simple as it may seem. Sometimes the objectives of an educational procedure are couched in such global terms (e.g., "the attainment of skill in arithmetic," "ability in creative problem solving") that it is not immediately possible to develop an evaluative procedure. The designer of the educational procedure

or product may have developed it without a clear and specific notion of his objectives; in which case it may be necessary to press him to make those objectives explicit before evaluation can begin. Frequently the effort to state objectives reveals a need to recast the educational procedure or product itself. Ideally, a statement of educational objectives includes specifications of detailed instructional content that the learner is expected to master, and specifications of the kinds of behaviors or performances that will, hopefully, certify the desired degree of mastery. When such statements are available, the process of translating them into evaluative instruments is facilitated, although it is never really easy.

Educational research workers find it useful, in formulating statements of educational goals, to make reference to a "taxonomy" of educational objectives such as that for the "cognitive domain" by Bloom (1956), or that for the "affective domain" by Krathwohl, et al. (1964). Bloom's taxonomy classifies objectives in the cognitive domain into the following broad categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation; each category contains a number of sub-categories. Bloom illustrates how these classifications can be represented by behaviors or performances that can be, within certain limits, incorporated into evaluative instruments.

It is usually helpful, also, to organize specifications of educational goals in the form of a two-way table in which the rows are labeled in terms of components of instructional content, and the columns represent kinds of behaviors (such as recognition, recall, problem solving, application to concrete situations) which will reveal mastery of that content. In filling out a table of this sort one is forced to decide upon the particular kinds

of objectives for which one desires evaluation, and then to choose or select adequate samples of goal specifications upon which to base evaluation instruments. One also becomes aware of objectives that may be more than usually difficult to use as bases for evaluation, and that may, in consequence, be left out of account unless special pains are taken.

The Construction of Evaluative Instruments

There is both science and art in the construction of evaluative instruments, whether they be objective multiple-choice tests, essay examinations, rating scales, performance tests, standardized interviews, or systematic observations of behavior in natural situations. (In this paper we use the term test in a generic sense to denote a wide variety of measuring procedures, any of which may play a role in an evaluative program.) The scientific aspects involved are in the realm of such matters as item sampling, item analysis, the assembly of item composites into tests, and the assessment of the reliability and validity of the measuring instrument. A large part of test theory, in fact, concerns problems having to do with how best to assemble a composite of separate test items in order to yield a measurement instrument with desired characteristics of reliability and validity. But there are other aspects of test construction that require perceptive intelligence and creative imagination on the part of the test constructor--relatively rare qualities. In general, there is no way of constructing an evaluative instrument "by formula," even though certain aspects of test construction may be done by a computer. The construction of a test requires as much creative ability as the writing of, say, an essay--but a different kind of ability, one that involves insight not only into the subject matter

(if it is a test of subject-matter mastery), but also into how that subject matter is perceived and learned (or can be misperceived and learned wrongly) by pupils. For example, in constructing multiple-choice questions the item-writer must not only be able to state clear questions but also be artful in proposing "distractors" (wrong alternative answers) that will be plausible to the student with limited knowledge and yet not attractive to the student with adequate knowledge. The work of the item-writer is to some extent controlled by the statistical results obtained with his items, as when statistical analysis discloses that an item does not adequately discriminate well between students possessing adequate knowledge and those who have only partial knowledge, or less. However, statistical analysis is no substitute for the perceptiveness and creative ability of the test constructor.

Certain types of educational objectives are easier to test than others. It is relatively easy to test for the presence of factual knowledge or elementary skills in such subjects as science and mathematics; it is more difficult to assess a pupil's creative writing ability, ability to speak a foreign language, "inventiveness" in mathematical problem solving, or grasp of major historical trends. Partly the difficulties are semantic--the objectives may be difficult to define in the first place; partly difficulties are practical and can be overcome only by unusual arrangements or efforts. Early examples of unusual yet ingenious and feasible procedures for measuring certain "difficult to measure" educational goals are to be found in the work of Hartshorne and May (1928), on the assessment of such character traits as honesty. Often, relatively simple evaluative devices can be found which measure certain objectives somewhat indirectly and yet

validly. For example, certain kinds of objective tests of ability to discriminate good and poor writing have been found to be highly correlated with more elaborate tests of creative writing ability, and hence, for some purposes may be used as reasonable adequate substitutes for the latter.

It should be emphasized, in any case, that the development of satisfactory evaluative instruments often requires much effort, imagination, and technical sophistication. The evaluative instruments themselves must be evaluated. There is no guarantee, further, that in any particular instance a satisfactory evaluative instrument can be developed; some educational objectives seem to be essentially unmeasurable.

Evaluative instruments vary in the extent to which they are an integral part of the instructional process. The traditional practice has been to intersperse evaluative procedures in the course of instruction, e.g. a test at the end of every unit. Sometimes evaluations are completely external, as when a standardized test is given to a group of students under the auspices of an outside agency like the College Entrance Examination Board. At the other extreme, evaluation is built into the instructional process itself, as where a teacher uses a "Socratic" method to develop knowledge and insight in the pupils; similarly, "programmed instruction," whether purveyed by "programmed textbooks" or a computer console, characteristically proceeds by asking students questions covering the material presented, student progress often being contingent upon his successful response to these questions. In some types of programmed and/or computer-based instruction, the student may be "branched" to more advanced material if he is more successful than the average student, or he may be shifted to special remedial material if he has more than ordinary difficulty with the main-line program. This

"branching" action of the program, if it is to be effective, depends upon the presence of appropriate diagnostic and evaluative features in the program itself. Thus, at least in situations where the prior planning and control of instructional procedures with built-in evaluative features is possible, the principles and findings of educational measurement can be usefully applied. (In fact, the problems posed by built-in evaluative procedures require special extensions of classical test theory.)

Enter Technology

In trying to propose a role for measurement and evaluation in "educational technology" I feel a need to state what I shall mean by this phrase. "Technology" is a relativistic term; it can pertain to any device or procedure which makes use of scientific knowledge. I have already indicated that educational measurement is itself a technology. Further, the very process of instruction can be regarded as a technology, to the extent that it is based on a theory of instruction. One kind of educational technology, for example, is "programmed instruction," which is based on a set of principles derived from psychological theory and which can be conducted with the simplest of materials or devices, e.g. the "programmed textbook." Yet it must be included in any definition of educational technology. One's ordinary associations with the phrase prompt one think, however, of specialized machines or devices that are based on contemporary industrial technology and that are, or can be, used in educational settings for presenting, recording, or otherwise processing information of a visual or auditory character--devices such as the film projector, the television receiver, the tape recorder, and (above all) the modern computer. I say "above all" the modern computer because it can control an assemblage of

other devices and can even supplant some of these other devices. We shall consider the role of measurement and evaluation in connection not only with programmed instruction but also with technological devices for presenting, recording, or otherwise processing information.

Three trends are seen in the development of educational technology:

(1) More efficient and flexible ways of presenting stimulus material (e.g., random access to a file of material to be presented visually), or of recording visual and auditory information.

(2) Increasing control and monitoring of the interaction between the student and the stimulus material (e.g., with a computer, capability whereby the student can respond to the stimulus with a light pen in such a way that the computer senses and records the response and takes further action contingent upon this response).

(3) Increasing capability for complex processing of data from student responses.

Trend (1) has long been evident in the development of such devices as the phonograph, radio, film, and TV. Trends (2) and (3) have been more fully realized only with the advent of the computer.

Trend (1)--more efficient and flexible ways of presenting stimulus material--has aided educational measurement and evaluation in numerous ways. For example, the invention of the tape recorder made it more convenient to present auditory stimulus material in connection with certain kinds of tests. A number of school systems use their own radio or TV installations regularly to administer school-wide tests and examinations: such a procedure standardizes the conditions of test administration. Further, recording devices

such as the videotape recorder have facilitated the storing of classroom observations and records of teacher performance for later evaluative analysis.

The Evaluation of Presentation Devices.

As used in the conduct of instruction, technological stimulus presentation devices such as the phonograph, movie film, or TV are only as good as the material that is presented through them. Sometimes they have added advantages such as greater convenience, richer possibilities with respect to the variety of material presented, and greater interest and better attitudes on the part of the students, but these bonuses do not automatically accompany these technological devices. Student attitudes, for example, have been found to be partly dependent on the attitudes of their teachers towards the technological device, or upon the quality of the material presented. Devices that do nothing but present materials are likely to have certain limitations as instructional media: usually they do not allow self-pacing by the student or variations in the material presented to the various students in a class. It may be inconvenient for the student to take notes on the material, and the possibilities for immediate response and feedback are often quite small.

Most research studies attempting to evaluate the use of film or television have found "no significant differences" between the results of such use and those of more traditional methods of instruction (Allen, 1960; Reid and MacLennan, 1967; Lumsdaine, 1963). This is only a generalization, however, there are studies which have indicated ways in which films and TV presentations can be improved and used more effectively. Even the re-showing of a film can improve learning markedly. Further, even if there are no large differences between the use of films and TV and the use of

more traditional methods, it will often be the case that the educator can confidently supplant traditional instruction by introducing newer media, with consequent economic benefits such as the conservation of teacher manpower.

In nearly all the research studies on the evaluation of newer media, educational measurement has played a large role in measuring the characteristics of pupils or classes at various points in the course of instruction--before instruction begins, during instruction, and at the end of instruction. Student achievements are measured by standardized or special-purpose tests, and their attitudes are assessed by various types of attitude scales constructed according to psychometric principles. Nevertheless, several criticisms can be made of these evaluative studies:

(1) The design of the studies often leaves much to be desired. (In one review of research [Stickell, 1963] it was claimed that of 250 comparisons between televised and face-to-face instruction, 217 were classified as "uninterpretable" because of poor research design.)

(2) The measures of student achievement are sometimes of poor psychometric quality, with low reliability and/or validity, insufficient attention being given to the construction of proper evaluative instruments. One of the most frequent errors is the failure to make certain that the achievement tests that are constructed cannot be passed by individuals who have <u>not</u> had the instruction being investigated. Otherwise, test items can frequently be passed by individuals on the basis of general intelligence or general information rather than on the basis of specific instruction.

(3) The studies are nearly always of the "summative" variety; very few attempt to find particular defects in the instructional material or its use and correct those defects by "formative" evaluation. One exception is the study of Gropper and Lumsdaine (1961) who used student responses (errors on test items) to make successive improvements in a kinescope--improvements that paid off in significantly better student performance. If more "formative" evaluation were done for materials presented by film or television, the advantages of such presentations would probably be much enhanced. Unfortunately, people seem to resist the idea of editing films and kinescopes, once they have been brought to production standards.

Measurement and Evaluation as Related to Programmed Instruction

Programmed instruction has three distinguishing characteristics: (1) It is based on a detailed analysis of educational objectives, the objectives being stated in "behavioral" terms; (2) the steps of the instruction ("frames") are carefully chosen, sequenced, and organized--usually they are relatively "small" steps where the student's attention is directed to only one or a very small number of newly-presented elements to be learned at a time; (3) the program is normally arranged so that the student receives immediate confirmation of correct responses. Most programs are intended to be given to students under self-pacing conditions. A special kind of "formative" evaluation is employed in the development of the better programs: programs are tried out on small samples of students to detect errors and are then successively revised until error rates are low. As noted earlier, testing materials are usually built directly into the program, both in the form of "prompted" teaching frames and in the form of "unprompted" frames in which the student has to demonstrate

mastery without the presence of cues or other helps. Some programs also present, at the end, a final test of a fairly conventional character.

Because the object of programmed instruction is to produce complete or nearly complete mastery, it has sometimes been argued that conventional principles of item analysis do not apply to the testing materials built into programmed instruction or even to "summative" evaluation materials given after the student has completed a program. Conventional principles call for items that are passed by, say, 10% to 90% of the sample, whereas programmed instruction tests should be passed by 100% of the sample. This argument ignores the fact that even in the context of programmed instruction, test items must be reliable and valid indicators of something, namely mastery of the skills or knowledges which are hopefully taught by the program. Thus, they should discriminate between pupils who have learned through the program and pupils who have not had the program or its equivalent. The test represented by an unprompted frame should, indeed, be passed by 100% of pupils going through the program, but it should be passed by a significantly lower percentage of pupils who have not had the program or equivalent instruction. Holland (1965) has shown that many frames in poorly constructed programs do not really teach or test, because he finds that even when large portions of the material in the frame are deleted ("blacked out") the pupil can still give the desired response; his "black-out" technique, he claims, provides a measure of the degree to which the material is properly programmed. Holland's technique is thus a logical extension of traditional concepts of test construction, since he shows, in effect, that certain test frames in

instructional programs do not discriminate between those who have mastered the material and those who have not.

In appraising "programmed instruction," that is, in applying "summative" evaluation to programs, workers in the field have tended to eschew attempts to compare the effectiveness of programmed instruction with that of other kinds of instruction. They are more concerned with demonstrating the effectiveness of this kind of instruction in terms of its own goals. They insist that properly prepared programs should be accompanied by detailed information as to (1) the kinds of learners for which the program has been designed and validated, and (2) the achievement attained by those learners (in terms of time to reach criterion performance, error rate, or performance on criterion tests). One definition of a "program" has it that it is "a vehicle which generates an essentially reproducible sequence of instructional events and accepts responsibility for efficiently accomplishing a specified change from a given range of initial competences or behavioral tendencies to a specified terminal range of competences or behavioral tendencies" (Lumsdaine, 1964, p. 385). The acceptance of this responsibility, on the part of a program writer, entails the responsibility to provide the necessary proof of effectiveness; that proof will often be supported by evidence from before-and-after tests and other observations of performance.

One can, of course, use standard experimental designs to compare the effectiveness of "programmed instruction" with other types of instruction, including traditional classroom instruction. In the relatively few comparisons of this type, programmed instruction has come off rather well (Schramm, 1964), often because it affords a more efficient approach to instruction in terms of time taken to learn and amount retained after a

lapse of time. It remains true, of course, that there are both good and poor programs just as there are good and poor teachers. Therefore it is difficult to make any generalizations, and perhaps one should not attempt to make them, except to say that programmed instruction, like any other form of special instruction, merits careful consideration for regular use in schools. Although programmed instruction has not been the panacea that it was first thought to be, it seems to have attained a solid place in educational programs and may even increase in acceptance, as better programs are prepared. Its popularity in industrial and governmental training programs is a testimony to its usefulness.

Enter the Computer

In the above discussion of programmed instruction we might have mentioned the teaching machine, i.e., any device for presenting the materials of instruction and arranging for the correctness of student response to be confirmed or disconfirmed. In fact, simple teaching machines were developed as early as 1915 by Pressey, and Skinner's early work in programmed instruction, around 1954, included construction of several teaching machines. There has been some rather inconclusive research on whether use of teaching machines yields greater effectiveness than the use of printed materials like the programmed textbook. The machines used in this research were often somewhat unreliable, inconvenient, and too expensive. Further, most of them were relatively simple, being limited to systematic, sequential presentations with confirmations or disconfirmations of student response. Today, more reliable and complex teaching machines are available, but there has been little research to evaluate them.

The advent of the modern computer in educational settings, around the middle 1950's, brought a new realm of possibilities, including increased complexity by several orders of magnitude. The first computers, like the early teaching machines, were somewhat unreliable and expensive, but at this writing we are going into the fourth generation of computers--even more expensive than before, but fast and powerful enough, it would appear, to reduce the cost of the student instructional hour to a small figure, perhaps something like 25 cents (according to one recent estimate) even taking into account the costs of program development, author royalties, remote communication lines, etc. This figure is competitive with ordinary classroom instruction. After a period of frank skepticism, I have become convinced that the computer will play an increasing role in instruction at all educational levels, and therefore I feel justified in giving it special attention in this paper.

What gives the computer its special promise is that it makes possible, much more than noncomputerized "teaching machines," the development of the second and third technological trends mentioned above, namely, increased control and monitoring of the interaction of the student with the stimulus material, and increased capability for complex processing of data from student responses. With respect to the former trend, the computer can orchestrate a whole panoply of other devices (such as film display units, sound-track storage-and-display mechanisms, TV monitors, and special student response devices) along with the by now conventional teletype keyset. With respect to the latter trend, it may be noted the computer can not only store and analyze multitudinous data about student responses (speed of response, correctness, freely composed answers, etc), but it can also utilize complex

logic in making well-nigh instantaneous decisions about those responses and what is next to be presented to the student.

The almost unlimited capabilities of the computer enable it to be used in a wide variety of educational settings, at all educational levels. It can even simulate, in a realistic way, a free dialogue between student and tutor, so long as the student is able to type his responses on a keyset. One of the obvious limitations of the computer (at least in terms of presently-available technology) is that it is largely limited to the exchange of alphanumerical information with the student, and to the presentation (not the reception and evaluation) of visual and auditory material. It cannot evaluate students' oral responses or motor performances unless those can be translated into the digital input required by the computer, and successfully evaluated by the computer logic.

There are numbers of ways in which the computer can be used in instruction; in "computer-assisted" instruction the student is "on-line" with the computer and stored in the computer configuration (Stolurow & Davis, 1965; Atkinson & Wilson, 1968); in "computer-managed" instruction, the computer helps the teacher to administer and guide the instructional process, but the student is not "on-line" with the computer (Brudner, 1968).

In computer-assisted instruction as it has developed to date, many of the principles developed in programmed instruction are applied: careful analysis of educational objectives, development of programs by tryout and revision, use of relatively small steps in the instructional presentation, use of immediate feedback to confirm the student's responses. What we have said about the application of educational measurement and evaluation to programmed instruction also applies, in large measure, to computer-assisted

instruction. That is to say, the evaluative process is usually built into the program "software" that is operated by the computer, and the effectiveness of the system is judged in terms of the speed and efficiency with which students attain the stated instructional objectives.

As yet there are few studies comparing computer-assisted instruction with other forms of instruction. Experiences with computer-based instruction in reading, arithmetic, and Russian at Stanford University indicate that learning (as measured by standardized or special-purpose tests) is at least as efficient as under more traditional instruction. In the case of Russian, there were fewer drop-outs from the computer course than from the conventional classroom. It is likely that research of the "comparative effectiveness" type will yield the same kinds of conclusions as other kinds of comparative effectiveness studies--that in general there are "no significant differences" in attainment, and that attainment is a function, not of the machine itself but of the quality of the instruction, however conducted--that is, the way in which the instructional content is put together, tried out, revised, and validated.

Lest the above paragraph give too pessimistic an impression, however, I hasten to say that I believe the computer will in time render an enormous service to education. It will make it possible to offer more different courses to more students, and to guarantee student attainment to an extent not previously thought possible. This will come about, at least in part, through the intelligent application of principles of educational measurement and evaluation. To be specific:

(1) Because of its capability for storing and analyzing student responses, the computer will facilitate the "item analysis" of instructional content and

the tryout and revision of instructional programs. Already at the University of Illinois, it is standard procedure to print out daily error analyses for computer-course authors, who then try to revise their programs to reduce student error.

(2) The computer is an enormously convenient testing device. It can in the first place rather quickly diagnose the student's initial state of knowledge about a subject-matter, "branching" him either to easy or difficult material according to his needs. In the second place, it administers quantities of test materials in the course of an instructional program; the student is not allowed to progress through the program unless he demonstrates mastery at intermediate points. Three, it can easily administer most standardized tests, quickly producing not only the conventional raw score but also diagnostic information on particular types of difficulties, information on speed and correctness of response to particular items, plain-language interpretations of test scores, and the like. Use of consoles at remote locations might make possible the administration of standardized tests simultaneously over wide geographical areas--even computerized nationwide test administration (as of College Board tests) is not out of the question.

(3) The computer can accumulate and analyze data on large numbers of students--data on student characteristics, learning performance, backgrounds, etc. It would thus enormously facilitate the evaluation of different instructional programs and the tabulation of the results. Whether or not it is used in computer-based instruction, it could accumulate large amounts of readily-analyzable information on the total educational program that could be provided to educational researchers and administrators in easily

comprehensible form. Already this sort of thing is done in the state of Iowa in the public education system.

(4) Specialized capabilities may be developed whereby computers can evaluate free responses of students as validly, and more efficiently, than they can be evaluated by teachers. Work is now going on at the University of Texas whereby students' answers to essay questions in science courses can be quite accurately scored by computer. Ellis Page, at the University of Connecticut, is working on programs to grade high-school students' English compositions by computer, to diagnose their difficulties, and provide remedial instruction (Page, 1966).

(5) The computer can also be used for various types of content analysis of instructional material. For example, work is now progressing on automating the process of measuring the "readability" of prose; readability (reading ease or comprehensibility) has been found to be an important variable in the effectiveness of textual materials. It may also be suggested (although this does not exactly fall within the purview of this paper) that computers may perhaps be programmed to <u>generate</u> instructional programs or at least certain components thereof.

<u>Summary</u>

Educational measurement and evaluation is itself a technology which is central to the operation and improvement of the educational process, because it enables the educator to know crucially important things about pupil characteristics and achievements. It also furnishes him with a valid basis for judging the worth and effectiveness of educational programs and innovations, and improving them in both gross and detailed features.

There is a long history of the application of this technology to the development and evaluation of various educational innovations such as film, television, and "programmed instruction." At present, the computer is seen to be the important educational tool of the future. As in the case of other educational tools, the computer will be valuable only to the extent permitted by the quality of the instructional materials and programs put into it. Much research and development, using the technology of measurement and evaluation along with other technologies, will be necessary to allow the computer and other educational media to reach maximal usefulness in education.

There is still a large gap between what is possible to accomplish through measurement and evaluation and what has actually been accomplished. This gap can be filled by training more research and development specialists, training teachers and administrators to utilize research and development results more effectively, and providing adequate funds for these training, research, and development activities.

REFERENCES

Allen, W. H. Audio-visual communication. In Harris, C. W. (Ed.) Encyclopedia of educational research (Third Edition). New York: Macmillan, 1960. Pp. 115-137.

Anastasi, Anne. Differential psychology: individual and group differences in behavior. New York: Macmillan, 1958.

Atkinson, R. C., and Wilson, H. A. Computer-assisted instruction. Science, 1968, 162, 73-77.

Bloom, B. S. (Ed.) Taxonomy of educational objectives. Handbook I: Cognitive Domain. New York: Longmans, 1956.

Brudner, Harvey J. Computer-managed instruction. Science, 1968, 162, 970-976.

Buros, O. K. (Ed.) The sixth mental measurements yearbook. Highland Park, New Jersey: Gryphon Press, 1965.

Campbell, Donald T., and Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 171-246. [Also published as a separate by Rand McNally.]

Gage, N. L. (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963.

Gropper, G. L., & Lumsdaine, A. A. The use of student response to improve televised instruction: An overview. Report No. 7 (Summary of six prior reports), Studies in televised instruction. Pittsburgh: American Institutes for Research, Rept. No. AIR-C13-61-FR-245(VII), 1961.

Harris, C. W. (Ed.) Encyclopedia of educational research. 3rd edition. New York: Macmillan, 1960.

Hartshorne, H., and May, M. A. Studies in deceit. New York: Macmillan, 1928. 2 vols.

Holland, James G. Research on programing variables. In Glaser, Robert (Ed.) Teaching machines and programed learning. II. Data and directions. Washington, D. C.: Department of Audiovisual Instruction, National Education Association, 1965. Pp. 66-117.

Krathwohl, D. A., et al. Taxonomy of educational objectives. Handbook 2: Affective domain. New York: McKay, 1964.

Lord, F. M., and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lumsdaine, A. A. Instruments and media of instruction. In Gage, N. L. (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963. Pp. 583-682.

Lumsdaine, A. A. Educational technology, programed learning, and instructional science. In Hilgard, E. R. (Ed.) Theories of learning and instruction: Sixty-Third Yearbook. National Society for the Study of Education. Chicago: Univ. Chicago Press, 1964. Pp. 371-401.

Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In Glaser, Robert (Ed.) Teaching machines and programed learning. II. Data and directions. Washington, D. C.: Department of Audiovisual Instruction, National Education Association, 1965. Pp. 267-320.

Page, Ellis B. The imminence of grading essays by computer. Phi Delta Kappan, 1966, 47(5), 238-243.

Reid, J. C., and MacLennan, D. W. Research in instructional television and film: summaries of studies. Washington, D. C.: U. S. Government Printing Office (Catalog FS 5.234:34041), 1967.

Schramm, W. The research on programed instruction: an annotated bibliography. Washington, D. C.: U. S. Department of Health, Education, and Welfare, Office of Education, 1964. (Catalog No. FS 5.234:34034)

Scriven, M. The methodology of evaluation. In Tyler, R. W., et al. Perspectives of curriculum evaluation. Chicago: Rand McNally, 1967. (AERA Monograph Series on Curriculum Evaluation, I.) Pp. 39-83.

Stickell, D. W. A critical review of the methodology and results of research comparing televised and face-to-face instruction. Doctoral dissertation. The Pennsylvania State University, June 1963.

Stolurow, L. M., and Davis, D. Teaching machines and computer-based systems. In Glaser, Robert (Ed.) Teaching machines and programed learning. II. Data and directions. Washington, D. C.: Department of Audiovisual Instruction, National Education Association, 1965. Pp. 162-212.